Can Speech Recognizers Measure the Effectiveness of Encoding Algorithms for Digital Speech Transmission?

C. Michael Chernick, Stefan Leigh, Kevin L. Mills, and Robert Toense
National Institute of Standards and Technology
January 1999

ABSTRACT

Modern communication channels, such as digital cellular telephony, often convey human speech in a highly encoded form. Methods that rely on human subjects to evaluate the quality of such channels are too costly to deploy on a large scale; thus, automated methods are often used to model quality as perceived by humans. Traditional automated methods that use Signal to Noise Ratios (SNR) to judge the quality of channels do not model human perception well when applied to highly encoded speech. For this reason, researchers investigate alternative means to objectively measure the quality of such channels. In this paper we explore the feasibility and applicability of using automated speech recognition technology to model human perception of the quality of communication channels that carry highly encoded (compressed) human speech.

We selected segments of speech from a widely accepted speech data base, and passed those segments through a speech recognizer under three conditions: (1) without encoding, (2) with encoding and decoding using a standard algorithm for speech compression, and (3) with encoding, transmission across a noisy channel, and then decoding. Speech recognition scores were computed for each speech segment under each condition. We then selected a subset of the speech segments, and asked human listeners to subjectively evaluate the intelligibility of the speech under the same conditions earlier input to the speech recognizer. We computed the correlation between the intelligibility of speech as evaluated by the automated recognizer and the human listeners. For speech segments used to train the recognizer, the correlation was .816 ± .064 (2 stdevs). For other speech segments, the correlation was .745 ± .074 (2 stdevs). These results are sufficient to encourage us to investigate the performance of commercial speech recognizers against human listeners on a more objective basis. Specifically, we envision scoring speech recognizers and human listeners on identical speech-to-text transcription tasks, and then computing the correlation in performance. If the next phase of this research yields acceptable results, then construction of an automated evaluation system, based on speech-to-text transcription, should be straightforward. Availability of an effective automated evaluation system will be useful to researchers and product engineers who are working toward advances in speech encoding algorithms for wireless communication channels and for Internet channels.

Keywords: network metrology, automated speech quality measurement, digital speech encoding, wireless communications, Internet voice

1. Introduction

Toll quality speech is limited to frequencies less than 4 kHz. Applying the well-known Nyquist sampling criterion, digitizing voice over this frequency range requires at least 8000 samples per second. Telephony applications allocate 8 bits of μ-law encoded data to provide toll quality voice calls. Thus, 64 Kbps of bandwidth are required to transmit digital speech with toll quality. However, since the actual information content in speech can be carried in much lower bandwidth, researchers seek methods to encode toll quality speech using only a few kilobits per second. For example, one such technique is Code Excited Linear Prediction, commonly known as CELP [CELP91]. Using CELP, the bandwidth requirements for speech have been reduced to 4.8Kbps with only marginal reduction in the quality of the speech signal. With this great reduction in bandwidth requirements comes a greater sensitivity to noise in signal transmission. Consider that for CELP each bit represents a greater portion of the signal than is the case for digitized, uncompressed speech. When transmitted over copper or fiber optic lines, noise is usually sufficiently small to go unnoticed by the listener. However, in the quickly growing field of wireless communication (e.g., cellular and cordless phones) noise is a much more serious problem. In addition, for applications where voice is sent over Internets, packet loss due to congestion can adversely affect the transmission of digital speech by causing dropouts or failure to meet real-time constraints. These factors contribute to a continued interest in coding algorithms for digital speech.

Methods to measure the relative effectiveness of coding algorithms are necessary in order to compare competing approaches over a range of conditions. The most common method employs human listeners to grade perceived speech quality by assigning an opinion score from a subjective scale, typically consisting of five values from excellent to unsatisfactory [KOHL97, LI98]. While producing the desired comparisons, methods that depend on human subjects are too costly and time consuming to deploy on a large scale. For this reason, we seek new metrics for automatically evaluating the effectiveness of speech encoding algorithms. Such metrics must be objective, economical (in both time and money), and reflective of speech intelligibility as perceived by human listeners. This paper reports results from a preliminary investigation of the use of automated speech recognition technology as a means to evaluate coding algorithms for digital speech.

The paper is organized into seven sections. First, we discuss related work. Second, we present our motivation. Third, we describe our research methodology, and discuss the speech samples we used for the experiments. Fourth, we describe our experimental results with respect to both an automated speech recognizer and to human listeners. Fifth, we discuss the correlation between the performance of the speech recognizer and the perceptions of the human listeners. Sixth, we identify some future research related to our proposed evaluation method. Finally, we present our conclusions from the current experiments.

2. Related Work

Traditional automated systems for measuring transmission channel quality employ signal-to-noise ratio (SNR) or segmental SNR (SEGSNR), or a frequency variant of SEGSNR [QUAC88]. SNR and SEGSNR are traditional metrics used in electrical engineering and related applications. While easy to measure and useful to assess selected encoding schemes, metrics based on SNR do not by themselves indicate the potential loss in recognition of compressed digitally encoded human voice signals. Quackenbush, for example, reports that metrics based on SNR apply only to waveform coders, to subband coders, and to adaptive transfer coders, but not to vocoder-like systems, such as CELP [QUAC88]. In fact, Quackenbush measured the correlation between subjective human listener opinions and objective scores for SNR (.24), SEGSNR (.77), and a frequency-weighted variant of SEGSNR (.93) across a variety of waveform encoders. His results indicate that only the frequency variant of SEGSNR correlated well with human perceptions. Quackenbush also evaluated a wide range of objective measures that could possibly apply to vocoder-like systems. Most of these measures exhibited poor correlation with human perception. One single measure, spectral distance, achieved a correlation of .80, while two composite metrics scored .82 and .86.

Researchers continue to search for objective quality metrics that can be applied to vocoder-like systems. For example, Kubichek and others report results from investigating several such metrics proposed to the International Telecommunications Union (ITU) for standardization [KUBI91, KUBI92]. As with SNR metrics, the metrics discussed by Kubichek measure differences between specific characteristics in channel input and output signals. Kubichek reports correlation between subjective human perception and three, single objective metrics: Cepstral Distance (.95), Coherence Function (.91), and Information Index (.83). Kubichek also reports correlation results for four composite metrics that use Bayesian estimation to seek relationships between parameter values and objective quality. Each of the composite metrics is composed from different combinations of single, objective metrics. The measured correlation ranged from .88 to .99. The best result was obtained by combining Cepstral Distance (CD) with Information Index (II). Other researchers report substantially different results. Bayya and Vis, for example, investigated objective measures for speech quality in wireless communications under different noise, distortion, and processing, and found that SNR and spectral distance measures provided the best correlation (.72) with subjective human perception [BAYY96]. Lam, and others, investigated objective speech quality measures for analog cellular telephones, and report the correlation with human perception for several metrics: Mel spectral distance (.86), Bark spectral distance (.84), coherence function (.81), and information index (.81) [LAM96].

Other researchers investigate the possibility of measuring the quality of speech channels by transforming the channel input and output signals into an internal representation of the sound that a human would hear. Beerends and Stemerdink propose a perceptual speech-quality measure (PSQM) based on transforming a physical signal into a psycho-acoustical model that considers the masking behavior of the human auditory system [BEER94]. They report correlation as high as .99 between PSQM and subjective evaluation by human listeners. Hauenstein also investigated the effectiveness of an objective measure akin to PSQM, and found correlation with human perception that ranged from .77 to .95, depending on the voice-coding algorithm in use [HAUE98]. Hansen and Kollmeier evaluated an objective measure for speech quality based on a psycho-acoustical model, finding correlation with subjective perception that varied from .88 to .93 depending on the voice-coding algorithm and the network connection in use [HANS97]. Similarly, Petersen and others, considering a psycho-acoustical model composed from four individual signal characteristics, computed a correlation of .94 with subjective ratings of speech quality from 40 subjects using 21 different rating scales [PETE97]. The correlation for each characteristic, when considered alone, was lower, ranging from .79 (bass/nasal) to .93 (hissing/crackling).

Voran and Sholl evaluated a number of objective measures for speech quality applied to a range of coding schemes and channel error conditions [VORA95]. Measures based on SNR and CD were found to be unreliable predictors of human assessment (correlation values ranged between .34 and .74, with one data set yielding a value of .97). More reliable predictions were observed for Bark Spectral Distortion (correlation ranged from .68 to .93) and for two variants of PSQM (correlation ranged from .74 to .89). Voran and Sholl conclude that highly detailed perceptual transformations, as proposed for example by Beerends and Stemerdink, do not prove particularly beneficial as predictors of human assessments of speech quality. Instead, Voran and Sholl suggest that improved distance measures provide better predictors.

While most proposed objective measures compare differences in input and output signals, Jin and Kubichek propose a metric based on comparing a quantized version of the output signal with a quantized version of a high-quality, reference signal [JIN96]. Their premise is that distorted speech will not match the entries in the reference codebook. Thus, they define metrics based on the distance between quantized units in the output signal and quantized units in the codebook. They test their metric against human opinion over a range of data sets, finding correlation values that vary from .46 to .93.

3. Motivation

Most previous work on objective measures for speech quality seeks some easily measurable combination of parametric differences between channel input and output signals that can reliably predict how humans will perceive the quality of the output signal. In the vast majority of cases, subjective human perception is captured as a mean opinion score (MOS) that ranges from 5 (excellent) to 1 (unsatisfactory) [KOHL97, LI98]. As discussed by Kubichek, the inherent variability of listeners and differing interpretations of the rating scale inhibits the reliability of MOS estimates [KUBI91]. These difficulties might compound as the number and granularity of scales to be scored increases. For example, Quackenbush measures subjective human perception for sixteen specific signal characteristics, where each characteristic can be distinguished on a 100-point scale [QUAC88]. This inherent variability might account for much of the variability reported by researchers who attempt to correlate objective measures with MOS.

To assess speech quality using objective measures, an automated measurement system should have five characteristics. First, the objective measures should provide a reasonably close approximation to human perception of intelligibility, and should be able to distinguish between degrees of intelligibility with the same resolution as human listeners. Second, tests should be repeatable, so independent tests of the same systems under the same conditions should achieve the same results. Third, the measurement system should operate effectively over a useful range of speech quality. Fourth, measurements must be computed within a reasonable delay. Fifth, the costs of making measurements must not be prohibitive.

In thinking about these characteristics, we investigated test methods for assessing the quality of speech recognition systems. We found a test method that uses reference speech data and performance metrics to compare the performance of various speech recognizers [HTK97, GARF93]. Inverting this method, we wondered if speech recognizers might be an effective reference against which to measure speech quality on digital transmission channels. Could changes in the performance of a speech recognizer reflect changes in the quality of speech as perceived by human listeners? If so, could an automated measurement system based on the objective performance of a speech recognizer attain the desired characteristics discussed in the preceding paragraph? We suspect a positive answer to both questions.

We foresee a quality score that might prove more reliable than MOS, yet still reflect differences in intelligibility as perceived by human listeners. Specifically, if a human subject were asked to transcribe the words from a speech segment and that transcription could be compared with the transcription generated by a speech recognizer for the same segment, then the correlation between human perception and automated objective measures could perhaps be made with greater reliability. If our ideas can be confirmed, then a new approach to objective measures for speech quality might prove feasible. Before proceeding to test our hypothesis, we decided to investigate how well a speech recognizer would perform as a predictor of MOS. This paper reports the findings of our initial investigation.

4. Research Methodology

Figure 1 illustrates the method used to generate speech samples for input to a speech recognizer, and for evaluation by human listeners. As documented in Table 1, we selected nineteen speakers from the TIMIT database, a widely accepted database of labeled speech segments that has been used to evaluate speech recognizers, developed by researchers with funding from DARPA (Defense Advanced Research Projects Agency). [GARF93] Although the TIMIT database includes speakers from eight different dialect regions within the United States, the speakers we selected were from only two regions: New England and Northern. We selected the New England region because we had access to a speech recognizer already trained for speakers from that region; this saved us time. We selected speakers from the northern region for variety. Each speaker spoke sentences, chosen for phoneme variety, ranging between 24 and 42 seconds in duration, with a median of approximately 30 seconds. In reporting results later in this paper, a unique number, taken from the TIMIT database, identifies each speaker. While the TIMIT database divides the speech samples from each region into two categories, training samples and testing samples, we used only the testing samples from the New England region (speakers number one through eleven in Table 1) to train the speech recognizer. This reinforced the recognizer’s initial training for speakers from New England.

Table 1. Information Describing Speakers Selected from the TIMIT Database

Speaker Number	TIMIT ID	Speaker Gender	Dialect Region	TIMIT Use	Recording Date	Birth Date	Educa-tion	Listening Tests
1	AKS0	Female	NE	Testing	04/01/1986	12/05/1956	PHD	USED
2	DAC1	Female	NE	Testing	03/12/1986	11/05/1918	HS	
3	ELC0	Female	NE	Testing	02/26/1986	05/20/1954	BS	USED
4	JEM0	Female	NE	Testing	02/11/1986	11/06/1951	MS	USED
5	DAB0	Male	NE	Testing	04/21/1986	12/15/1962	BS	
6	JSW0	Male	NE	Testing	04/16/1986	07/15/1953	MS	
7	REB0	Male	NE	Testing	02/25/1986	12/04/1958	BS	USED
8	RJO0	Male	NE	Testing	02/25/1986	05/15/1951	BS	
9	SJS1	Male	NE	Testing	02/11/1986	07/23/1960	BS	
10	STK0	Male	NE	Testing	04/17/1986	12/04/1960	BS	USED
11	WBT0	Male	NE	Testing	03/26/1986	05/24/1934	BS	USED
12	AEM0	Female	North	Training	02/27/1986	05/13/1960	BS	USED
13	DNC0	Female	North	Training	02/12/1986	04/27/1946	HS	USED
16	DLC2	Male	North	Training	04/09/1986	11/18/1959	MS	USED
17	DMT0	Male	North	Training	01/31/1986	07/13/1956	MS	USED
101	CJF0	Female	NE	Training	04/18/1986	08/31/1962	BS	USED
102	DAW0	Female	NE	Training	02/28/1986	07/18/1960	HS	USED
103	CPM0	Male	NE	Training	01/30/1986	02/18/1962	BS	USED
104	DAC0	Male	NE	Training	02/11/1986	11/29/1960	BS	USED

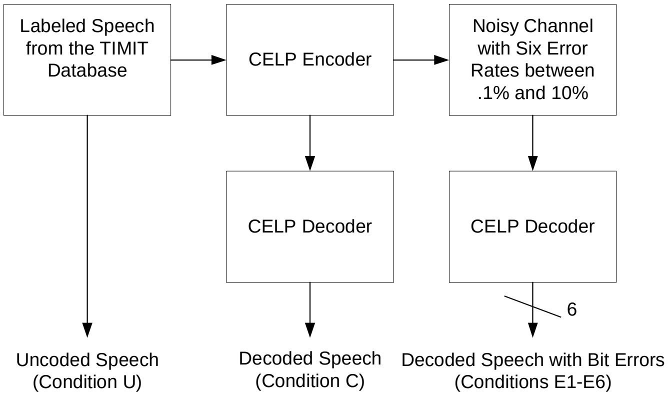
Labeled Speech from the TIMIT Database
CELP Encoder
Noisy Channel with Six Error Rates between .1% and 10%
CELP Decoder
CELP Decoder
6
Uncoded Speech (Condition U)
Decoded Speech (Condition C)
Decoded Speech with Bit Errors (Conditions E1-E6)

Figure 1. Method Used to Generate Speech Samples

As illustrated in Figure 1, phonetically labeled speech samples from TIMIT were processed along three different paths to produce digital speech for input to a speech recognizer and for evaluation by human listeners. The first path yields uncoded speech (Condition U) by simply forwarding the speech samples directly from the TIMIT database. The second path uses CELP to encode and then decode the TIMIT speech samples; thus, producing decoded speech (Condition C). On the third path, after CELP encoding, pseudo-random bit errors are introduced into the encoded data to simulate a noisy transmission channel, and then the speech samples are decoded. For the experiments reported here, we used six Bit Error Rates (BERs): .1%, .5%, 1%, 2%, 5%, and 10%, yielding six segments of speech (Conditions E1-E6) from each TIMIT sample. As a result of applying these three paths, the 19 speakers selected from the TIMIT database yielded 152 speech samples for input to the scoring phase of our experiment. In order to reduce the number of human listeners required for scoring, we selected only 14 of the 19 speakers from the TIMIT database. We elected to eliminate speakers with strong regional (New England) accents, which might have tended to confuse our listeners, who were primarily from the Mid-Atlantic region. Further, we discarded the samples produced with a BER of 10%, because we decided that all human listeners would judge these to be unintelligible. In summary, we scored the speech recognizer against 152 speech samples, while we asked human listeners to consider only 98 of those 152 samples.

Figure 2 depicts the general outline we used to score the speech samples, and then to assess the correlation between the recognizer results and human perceptions. The figure can be considered in three blocks: (1) automated scoring using a speech recognizer, (2) subjective scoring by human listeners, and (3) correlation analysis. We address each of these in turn.

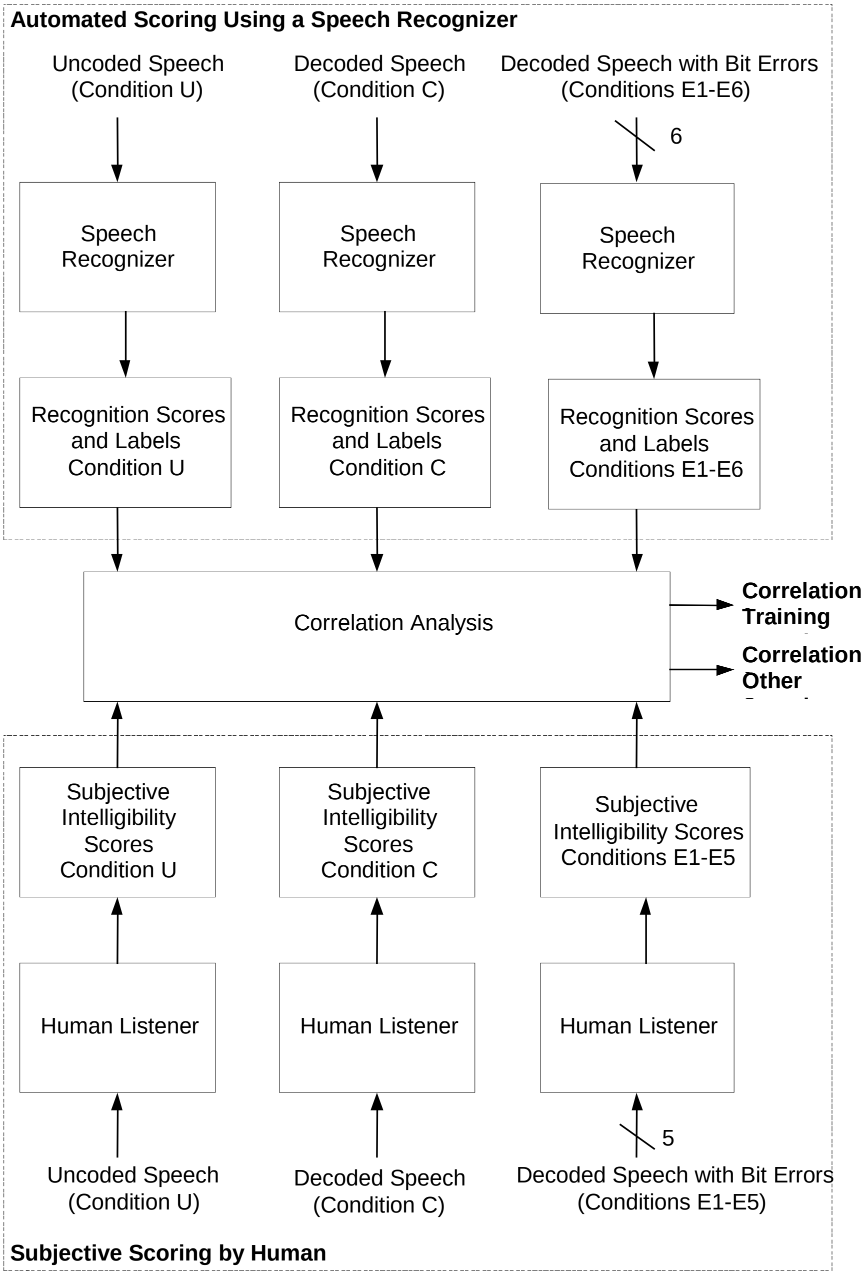
Automated Scoring Using a Speech Recognizer
Uncoded Speech (Condition U)
Decoded Speech (Condition C)
Decoded Speech with Bit Errors (Conditions E1-E6)
6
Speech Recognizer
Speech Recognizer
Speech Recognizer
Recognition Scores and Labels Condition U
Recognition Scores and Labels Condition C
Recognition Scores and Labels Conditions E1-E6
Correlation Analysis
Correlation Training
Correlation Other
Subjective Intelligibility Scores Condition U
Subjective Intelligibility Scores Condition C
Subjective Intelligibility Scores Conditions E1-E5
Human Listener
Human Listener
Human Listener
5
Uncoded Speech (Condition U)
Decoded Speech (Condition C)
Decoded Speech with Bit Errors (Conditions E1-E5)
Subjective Scoring by Human

Figure 2. Method of Scoring Speech Samples and Correlating Scores

Automated Scoring Using A Speech Recognizer. For the experiments reported here, we used a speech recognizer, HTK, readily available at NIST [HTK97]. The recognizer was included in the HTK Toolkit, available from Entropic Research Laboratory, Inc., and developed at Cambridge University. HTK uses Hidden Markov Models to achieve speech recognition [RABI93]. The HTK speech recognizer generates a set of speech labels indicating each phoneme recognized in an input speech sample. The recognizer adds a time stamp to each label. In order to score the performance of the recognizer, we used a scoring package included in the toolkit. The scoring package generates several statistics, including correctly recognized phonemes, insertions, deletions, and substitutions. In addition, the scoring package produces a confusion matrix indicating which phonemes were substituted for each of the input phonemes. In this experiment, we used only the number of phonemes correctly recognized, which we divided by the total number of phonemes in each speech sample in order to compute the percentage of phonemes recognized. Each of the speech samples was prepared for proper sampling rate and data format before being passed to the speech recognizer.

Subjective Scoring by Human Listeners. Since human listeners could not be asked realistically to identify phonemes in the speech samples, we had two options. We could ask human listeners to transcribe each speech sample, [EBEL95] or we could ask human listeners to subjectively score the intelligibility of each speech sample. In order to obtain a sufficient number of human listeners, we decided to reduce the time commitment for each listener by using subjective scoring.

We recruited fourteen volunteers to listen to and then subjectively score the intelligibility of speech samples played through a loudspeaker. Each volunteer was asked to listen to fourteen samples, and then to score each sample on a scale of one to five, as explained in Table 2, which gives the instructions for volunteers in the listening tests. We ensured that two different volunteers listened to each sample; thus, the 98 input samples were doubled to give 196 test samples.

Table 2. Instructions to Volunteers for Subjective Listening Tests

You will listen to a set of speech samples. Each sample consists of a single speaker saying several sentences. You will be asked to judge the overall quality of the several sentences. You may adjust the loudspeaker volume using the buttons on top of the loudspeaker. You will be given the opportunity to replay the speech sample before scoring.

Please judge the samples for intelligibility using the following scale:

5 – Excellent Quality and Easily Understandable
4 – Good Quality but still Easily Understandable
3 – Not Very Good Quality, but Mostly Understandable
2 – Poor Quality, Difficult to Understand
1 – Extremely Poor Quality, Mostly or Completely not Understandable

Note: Rightmost loudspeaker button raises volume.
Button next to rightmost button lowers volume.

Correlation Analysis. Before assessing the correlation between the speech recognizer and human listener results, we performed a visual screening to identify factors responsible for variation in the data. Given that we tested over a range of bit error rates and that we used two different classes of speech samples, those used to train the recognizer and those not, we expected that both of these factors would account for any variation observed in the data. We used a simple graphical analysis to verify our expectations. We did not analyze the human listener data in this manner because the low granularity of the responses as coded would yield little insight.

After assessing the experimental data, we used correlation analysis to estimate how well the speech recognizer scores predicted the judgment of human listeners. We considered separately two classes of data: data based on speech samples from speakers one through eleven (used to train the recognizer) and data based on the other eight speech samples. The results follow.

5. Experimental Results

The first experiment applied a speech recognizer to the various speech samples generated. The second experiment asked human listeners to evaluate a select, but substantial, subset of the speech samples generated.

5.1 Results from Automated Speech Recognition Tests

Table 3 gives the average phoneme recognition scores obtained for each experiment conducted using the speech recognizer against the generated speech samples. We used all 19 speakers, identified in column one. Each of the remaining columns gives the results for each speaker under specific encoding and error conditions. The second column depicts results without CELP encoding and decoding, and without introducing errors. The remaining columns show results with CELP encoding and decoding, and with a variety of bit error rates (BER) injected between the encoding and decoding steps. The injected bit error rates range from 0% to 10%. The shaded rows (speakers one through eleven) correspond to speakers used to train the recognizer. Note that for one sample, Speaker Number 6 with CELP .5% BER, the speech recognizer failed, so we report no data.

Table 3. Average Phoneme Recognition Scores from Speech Recognizer Experiments

Speaker	No CELP	CELP	CELP	CELP	CELP	CELP	CELP	CELP
Number	0% BER	0% BER	.1% BER	.5% BER	1% BER	2% BER	5% BER	10% BER
1	96.10	69.37	58.56	61.56	55.56	56.76	45.65	35.44
2	93.88	73.39	49.85	65.75	62.39	54.74	47.71	35.47
3	94.89	63.35	49.72	52.84	52.84	53.12	37.22	32.67
4	90.88	66.10	63.53	60.11	60.11	56.41	41.88	37.89
5	91.92	58.08	56.59	55.39	49.70	43.71	32.04	26.65
6	90.19	70.57	66.46	No Data	61.71	55.70	45.25	35.13
7	90.45	71.35	69.10	66.57	58.43	58.43	46.91	39.61
8	88.64	56.82	49.30	46.43	41.56	41.23	35.71	31.17
9	88.74	58.94	51.99	48.68	48.34	43.71	32.78	31.79
10	93.88	70.03	64.83	62.39	59.02	49.54	35.47	30.58
11	95.08	75.08	73.54	68.31	60.62	55.38	46.46	38.77
12	58.23	44.82	46.95	45.12	38.11	41.46	36.89	33.84
13	50.87	45.95	40.46	30.64	41.04	40.46	31.50	31.21
16	53.58	50.47	48.60	46.11	45.79	43.61	40.19	28.35
17	52.37	52.68	52.05	48.58	45.75	42.59	32.81	32.49
101	48.46	43.54	36.73	37.07	37.76	36.73	32.31	34.35
102	53.94	46.67	47.58	36.97	44.85	40.61	34.85	31.82
103	52.26	47.74	45.81	46.13	42.90	43.23	40.32	31.61
104	55.87	51.29	50.43	49.57	46.42	40.69	37.54	30.37

The box plots [TUKE77] of Figures 3, 4, and 5 provide a more easily comprehended view of the results. Each box plot gives the minimum, median, and maximum scores, and encloses within a bounding box the interquartile range, where the middle fifty percent of the scores fall. Figure 3 contains eight box plots used to visually screen the data. Each box plot in the figure graphs two distributions of data from one column of Table 3. One distribution, shown in gray, represents the performance of the speech recognizer on speech samples used to train the recognizer (speakers number 1 through 11), while the other distribution represents the performance of speech samples for the remaining speakers. These box plots confirm our expectation that the performance of the speech recognizer differs significantly for the training speakers versus other speakers across all error rates. For this reason, we chose to separate these two classes of speakers for purposes of computing correlation with the human listeners. Figure 4 shows the distribution of data obtained for each column of Table 3 across the eleven speakers used to train the recognizer. For each set of degradation conditions, the figure Similarly, Figure 5 portrays the results for the eight speakers who were not used to train the recognizer. A monotonically decreasing pattern is observed as expected.

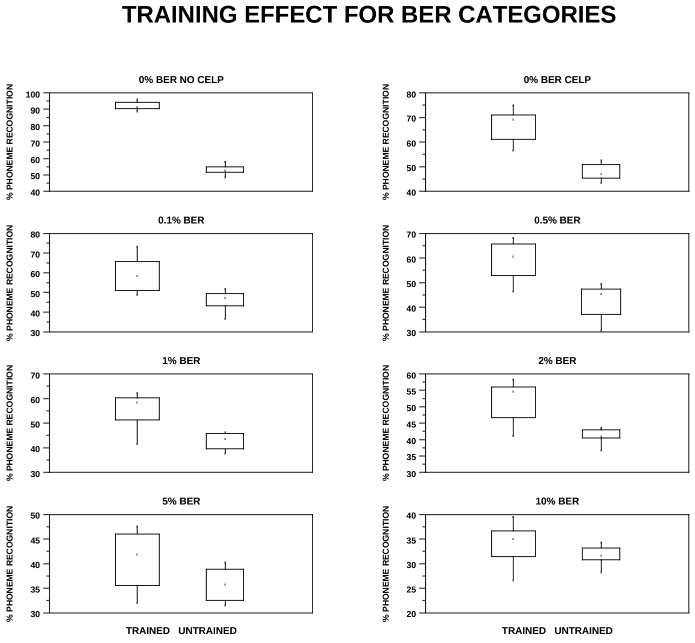

Figure 3. Visual Screening Showing the Training Effect for All Bit-Error Rate Categories

Some Observations. We were struck by the degree to which the speech recognition diminished for CELP-encoded speech samples, especially for speakers whose speech was used to train the recognizer. While not germane to the specific topic of this paper, we wonder how this result applies to the growing use of speech recognition technology across telephone calls. Will speech recognizers work as well for cellular telephones as they do for wire-line telephones?

5.2 Results from Human Listening Tests

Table 4 gives the average quality score computed for each speaker at each encoding and error rate in the human-listening tests. The last row of the table contains the average quality score computed across all speakers for the encoding and error rate depicted in each column. As expected, listeners judged the quality good to excellent for the unencoded speech and for CELP-encoded speech without errors. For CELP-encoded speech with .1% BER, the listeners found the speech understandable to good. As the BER reached .5%, listeners found the speech to be understandable but of poor listening quality. Understanding and quality dropped somewhat when the error rate rose to 1%. At the 2% error rate, listeners had difficulty understanding the speech samples. With a 5% BER listeners judged the speech to be practically unintelligible.

The reader should note that the human listening tests are much more subjective than the tests conducted with the speech recognizer because two listeners might judge the same speech samples differently. Also, the human listening tests have a much coarser, discrete scale for scoring the intelligibility of speech, as compared with the continuous percent phoneme recognition scale used with the speech recognizer. Despite these differences, we can still evaluate correlation between the results obtained with human listeners and those obtained with the speech recognizer.

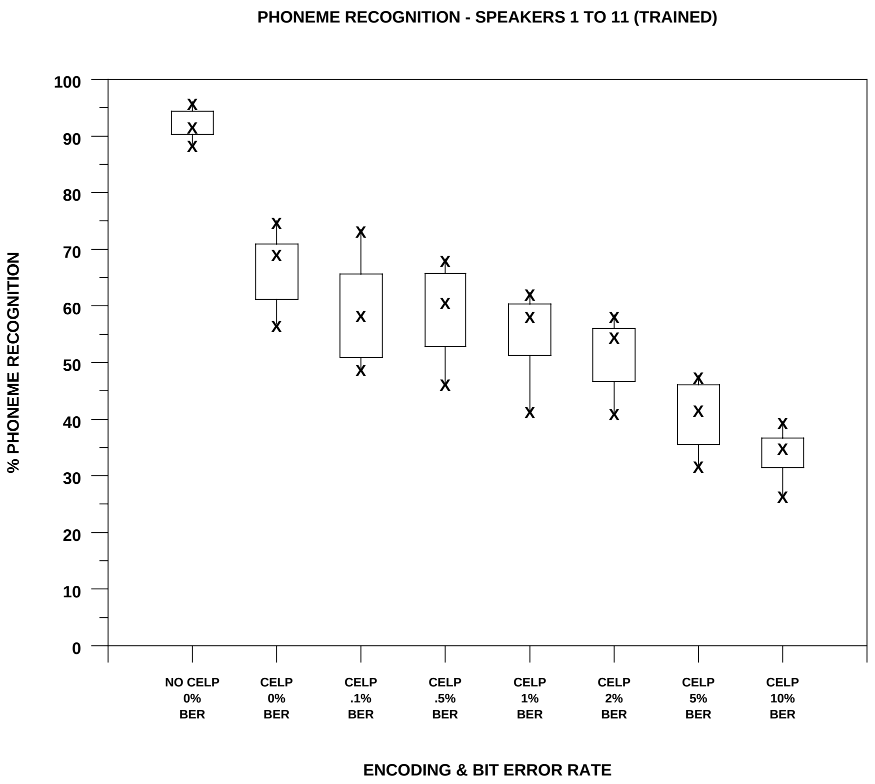

Figure 4. Performance of Speech Recognizer for Trained Speakers

Table 4. Average Speech Quality Assigned by Human Listeners

Speaker Number	No CELP 0% BER	CELP 0% BER	CELP .1% BER	CELP .5% BER	CELP 1% BER	CELP 2% BER	CELP 5% BER
1	5.0	4.0	4.0	3.5	3.0	2.5	2.5
3	5.0	3.0	3.0	3.5	3.0	1.5	1.5
4	4.5	3.5	4.0	3.0	2.5	3.0	2.0
7	5.0	4.0	3.5	2.5	3.0	2.0	1.0
10	5.0	4.0	3.5	3.5	3.0	2.0	1.0
11	5.0	4.5	4.0	3.0	4.0	2.0	1.5
12	4.5	4.0	3.5	3.0	2.5	2.5	1.0
13	5.0	5.0	3.5	3.0	3.0	2.5	1.5
16	4.5	4.0	4.0	3.5	3.5	2.0	1.0
17	5.0	4.0	3.5	4.0	3.0	2.5	1.5
101	4.5	3.5	3.0	2.0	2.0	1.5	1.0
102	5.0	3.5	4.0	3.0	2.5	2.0	1.0
103	4.5	3.5	3.0	3.0	1.5	1.5	1.0
104	5.0	5.0	4.0	3.0	2.5	2.0	1.0
All	4.82	3.96	3.61	3.11	2.79	2.11	1.32

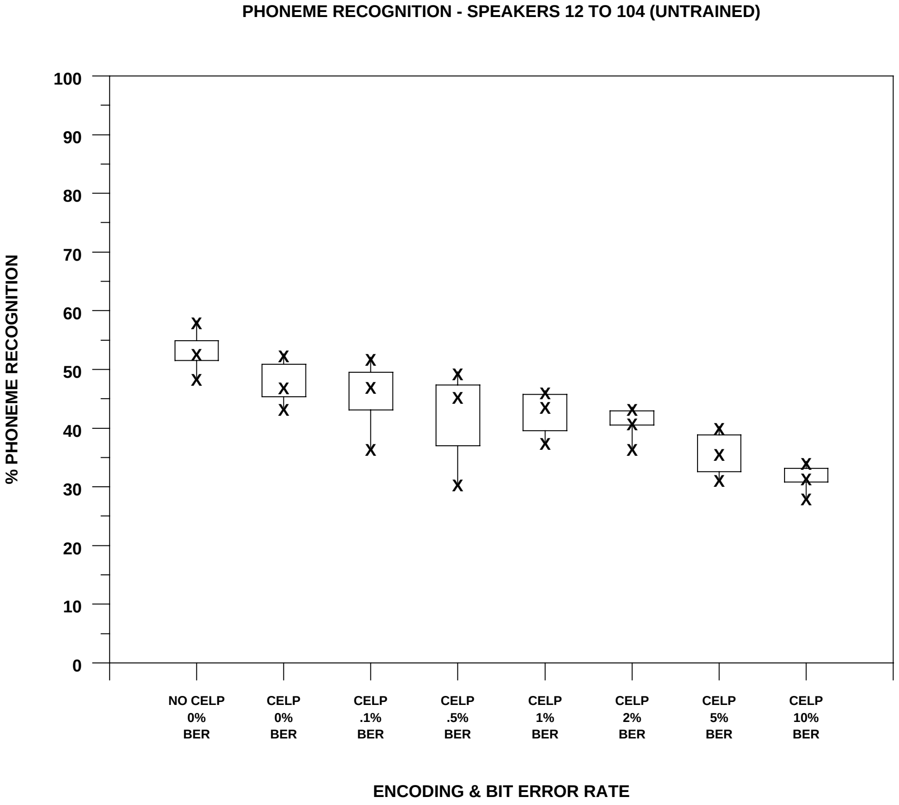

Figure 5. Performance of Speech Recognizer for Untrained Speakers

6. Results from Correlation Analyses

To assess the degree to which results from a speech recognizer can predict the quality and intelligibility of speech as perceived by human listeners, we compute the correlation between subsets of our two data sets. We consider the speech samples in two distinct sets: speakers used to train the speech recognizer and other speakers. Figure 6 plots human listener judgments against percent phoneme recognition from the speech recognizer for the eleven speakers used to train the speech recognizer. We computed the correlation at .816 ± .064 (2 stdevs). For the other eight speakers, Figure 7, we computed the correlation at .745 ± .074 (2 stdevs). The estimates of standard error were computed using resampling (bootstrap) [DIAC83]. To confirm these findings, we also computed correlation values using Spearman rank correlation [SACHS82]. The Spearman rank correlation for our training speakers is .789 and for the other speakers is .775. These values are the same whether the statistic is computed in the standard fashion as the correlation of the ranks of the scores, or whether an adjustment for ties in the scores is used. While the plots and correlation coefficients demonstrate clearly the monotone association between human and machine judgments of quality, the plots also display the inherent variation in the relation. Similar variation appears in other research the compares objective quality measures with human perception [BROO98].

7. Future Research

Our next step is to compare the ability of commercial speech recognizers and human listeners to transcribe speech samples under the same conditions reported in this paper. While we expect human listeners to be superior to speech recognizers in all cases, if we can establish a relationship between the performance of human listeners and speech recognizers, then we can consider building and deploying a test system for automatically scoring speech coding algorithms. We foresee a system that enables developers to select speech samples from a database and to select from among a range of speech recognizers. The developer could also select from a range of error models and rates, including independent bit errors, alternating periods of good and bad channel signals, and various packet switching network properties. With such a test system, developers could explore the properties of proposed speech coding and decoding algorithms under a range of conditions.

Beyond the use of speech recognizers for automated scoring of network-based speech coding algorithms, we can imagine applying techniques emerging from image understanding research to develop similar test systems for image and video coding schemes used for network transmission. Of course, image understanding research is less well developed than speech recognition research. Still, edge-detection techniques and object-extraction techniques seem worth investigating for this purpose. The development of multi-media coding and transmission algorithms could be greatly accelerated by the ability to automatically score performance in a manner consistent with human perception.

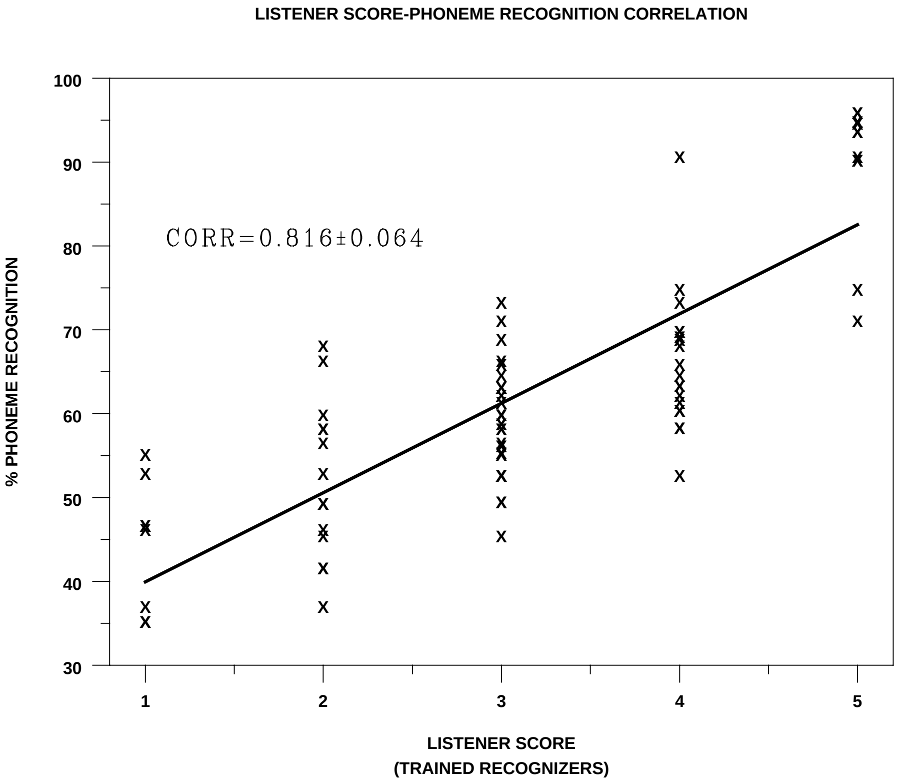

Figure 6. Correlation: Speech Recognizer and Human Listeners for Trained Speakers

8. Conclusions

We selected segments of speech from a widely accepted speech data base, and sent those segments through a speech recognizer under three conditions: (1) without encoding, (2) with encoding and decoding using a standard algorithm for speech compression, and (3) with encoding, transmission across a noisy channel, and then decoding. Speech recognition scores were computed for each speech segment under each condition. We then selected a subset of the speech segments, and asked human listeners to subjectively evaluate the intelligibility of the speech under the same conditions earlier input to the speech recognizer. We computed the correlation between the intelligibility of speech as evaluated by the automated recognizer and the human listeners. For unencoded speech segments used to train the recognizer, the correlation was .816 ± .064 (2 stdevs). For other unencoded speech segments, the correlation was .745 ± .074 (2 stdevs). Spearman rank correlation tests confirmed these numbers. These results are sufficient to encourage us to investigate the performance of commercial speech recognizers against human transcriptions. If the next phase of this research yields acceptable results, then construction of an automated evaluation system should be straightforward. Availability of an effective automated evaluation system will be useful to researchers and product engineers who are working toward advances in speech encoding algorithms for wireless communication channels and for Internet channels.

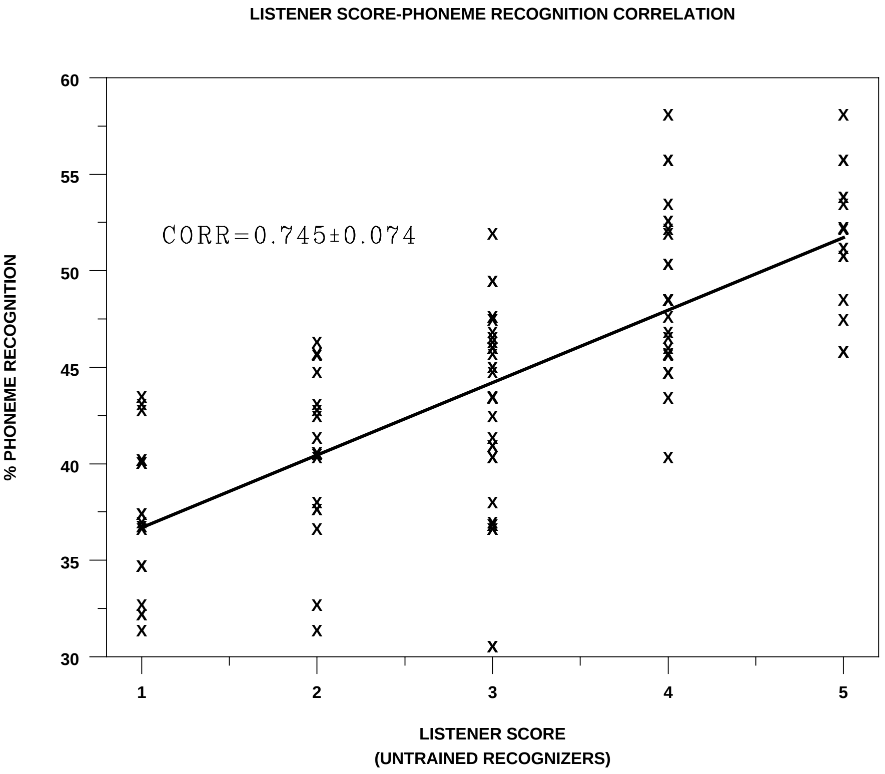

Figure 7. Correlation: Speech Recognizer and Human Listeners for Untrained Speakers

9. References

[BAYY96] "Objective Measures for Speech Quality Assessment in Wireless Communications", A. Bayya and M. Vis, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[BEER94] "A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation," J.G. Beerends and J.A. Stemerdink, *Journal of the Audio Engineering Society*, Vol. 42. No. 3, 1994.

[BROO98] "Getting the Message, Loud and Clear - Quantifying Call Clarity," S. Broom, P. Coackley, and P. Sheppard, *British Telecommunications Engineering*, Vol. 17, April 1998.

[CELP91] Federal Standard 1016, Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 Bit/Second Code Excited Linear Prediction (CELP), General Services Administration, Office of Information Resources Management, February 1991.

[DIAC83] "Computer-intensive Methods in Statistics", P. Diaconis and B. Efron, *Scientific American*, Vol. 248, pp. 116-130, 1983.

[EBEL95] Human Speech Recognition Performance on the 1994 CSR Spoke 10 Corpus, *Proceedings of the Spoken Language Systems Technology Workshop*, W.J. Ebel and Joseph Picone, January 22-25, 1995,

[GARF93] DARPA TIMIT Acoustic-Phonetic Continuous Corpus CD-ROM, John Garfolo, L.F. Lamel, William Fisher, John Fiscus, David Pallett, Nancy Dahlgren, NISTIR 4930, February 1993.

[HANS97] "Using a Quantitative Psycho-acoustical Signal Representation for Objective Speech Quality Measurement", M. Hansen and B. Kollmeier, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[HAUE98] "Application of Meddis' Hair-Cell Model to the Prediction of Subjective Speech Quality", M. Hauenstein, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998.

[HTK97] The HTK Book (for HTK Version 2.1), Steve Young, Julian Odell, Dave Ollason, Valthan Valtchev, Phil Woodland, Entropic Cambridge Research Laboratory Ltd., Compass House, 80-82 Newmarket Road, Cambridge CB5 8DZ, England, Tel: +44(0) 1223 302651 Fax: +44(0) 1223 324560, December 1997.

[JIN96] "Output-Based Objective Speech Quality Using Vector Quantization Techniques", C. Jin and R. Kubichek, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[KOHL97] "A Comparison of the New 2400 bps MELP Federal Standard with Other Standard Coders", M. A. Kohler, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997.

[KUBI91] "Advances in Objective Voice Quality Assessment," R. Kubichek, et al, *IEEE Global Telecommunications Conference*, 1991.

[KUBI92] "Advances in Objective Voice Quality Assessment," R. Kubichek, et al, *IEEE 42nd Vehicular Technology Conference*, 1992.

[LAM96] "Objective Speech Quality Measure for Cellular Phone", K.H. Lam, O.C. Au, C.C. Chan, K.F. Hui, and S.F. Lau, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[LI98] "Experimental Results on the Impact of Cell Delay Variation on Speech Quality in ATM Networks", B. Li and X.R. Cao, *Proceedings of the IEEE International Conference on Communications*, 1998.

[PETE97] "Objective Speech Quality Assessment of Compounded Digital Telecommunication Systems", K. T. Petersen, J. A. Sorensen, and S. D. Hansen, *Proceedings of the First Signal Processing Society Workshop on Multimedia Signal Processing*, 1997.

[QUAC88] Objective Measures of Speech Quality, Schuyler R. Quackenbush, Thomas P Barnwell, Mark A. Clements, Prentice-Hall, 1988.

[RABI93] Fundamentals of Speech Recognition, Lawrence Rabiner, Biing-Hwang Juang, Prentice-Hall, 1993.

[SACHS82] Applied Statistics: A Handbook Of Techniques, Lothar Sachs, Springer-Verlag, 1982.

[TUKE77] Exploratory Data Analysis, John W. Tukey, Addision-Wesley, Reading, Massachusetts, 1977.

[VORA95] "Perception-based Objective Estimators of Speech Quality," Stephen Voran, Connie Scholl, *Proceedings of the 1995 IEEE Workshop on Speech Coding for Telecommunications,* September 1995.